The Visual Computer
https://doi.org/10.1007/s00371-019-01721-7

ORIGINAL ARTICLE

Literature Explorer: effective retrieval of scientific documents through nonparametric thematic topic detection

Shaopeng Wu[1] · Youbing Zhao[1,6] · Farzad Parvinzamir[2,5] · Nikolaos Th. Ersotelos[4] · Hui Wei[1] · Feng Dong[1,3]

Abstract
Scientific researchers are facing a rapidly growing volume of literatures nowadays. While these publications offer rich and valuable information, the scale of the datasets makes it difficult for the researchers to manage and search for desired information efficiently. Literature Explorer is a new interactive visual analytics suite that facilitates the access to desired scientific literatures through mining and interactive visualisation. We propose a novel topic mining method that is able to uncover "thematic topics" from a scientific corpus. These thematic topics have an explicit semantic association to the research themes that are commonly used by human researchers in scientific fields, and hence are human interpretable. They also contribute to effective document retrieval. The visual analytics suite consists of a set of visual components that are closely coupled with the underlying thematic topic detection to support interactive document retrieval. The visual components are adequately integrated under the design rationale and goals. Evaluation results are given in both objective measurements and subjective terms through expert assessments. Comparisons are also made against the outcomes from the traditional topic modelling methods.

Keywords Topic explorer · Data visualisation · Topic modelling · Text mining · Web application · Scientific documents

1 Introduction

Owing to the growing volume of scientific publications, scientific document retrieval faces a high demand nowadays. It becomes increasingly difficult to manually gather and review a whole spectrum of publications in a target scientific domain over many years. To gain insight into a scientific field, scientific researchers need to explore the legacy publications in relevant topic areas in order to build a comprehensive picture. New approaches for advanced analytics are needed to help the researchers explore and retrieve the relevant documents of interest. More specifically, they require (1) easy access to the most relevant publications in their selected topics; (2) the knowledge about the relevant research topics and their relationships; and (3) the knowledge about the topic evolution over years.

Researchers often use thematic keywords that are commonly used in scientific fields to search for relevant papers. However, a naive search using only the user supplied keywords cannot yield satisfactory results, as there could be many documents containing these keywords and the word frequency alone does not provide a sufficiently good measurement on document relevance. To improve the retrieval outcomes, we need to discover sets of keywords that often co-occur in the same documents and are closely associated with the "thematic topics" that are generally recognised by the research communities. The thematic topics can be inferred from document collections. Similar ideas have been explored in the existing topic modelling approaches, in which each topic is defined as a collection of words [1, 9, 12]. In fact, topic modelling is an important research direction in natural language processing (NLP). Methods such as

✉ Shaopeng Wu
shaopeng.wu@gmail.com

Nikolaos Th. Ersotelos
n.ersotelos@wlv.ac.uk

1 University of Bedfordshire, Luton, Bedfordshire, UK

2 InsightZen LLC, Hangzhou, China

3 Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

4 Wolverhampton Cyber Research Institute (WCRI), School of Mathematics & Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton, West Midlands WV1 1LY, UK

5 Queens University Belfast, Belfast, UK

6 Communication University of Zhejiang, Hangzhou, China

Published online: 02 August 2019

latent Dirichlet allocation (LDA) [1] retrieve topic information by inferring hidden information according to the keyword distribution in documents. However, there are significant problems in retrieving scientific documents using the topics obtained by the existing topic modelling approaches. Many of them are duplicated, and there can be significant overlaps between different topics. In addition, some of these topic modelling methods create results that are not interpretable by human researchers. For example, latent semantic indexing (LSI) [2] generates negative values which are very difficult to understand; LSI [2] also finds topics that may not be explainable to humans. Other significant limitations include:

- While the nonparametric approach has the advantage to handle the large volume of scientific documents, many of the existing topic modelling methods are still parametric. For example, to run LDA, we need to specify the topic number, and to run hierarchical topic modelling, we need to define the levels of the topic hierarchy. All these settings are not practical to prefix in a scientific research area that comes with a highly dynamic nature.
- Many topic modelling methods face data scalability challenges. Methods like LDA are known for low speed performance owing to the number of iterations involved in the computation. Also many topic modelling methods generate thousands of topics in their applications in a large-scale corpus. Exploring scientific documents presents a very high demand in performance speed as the users need to make a variety of queries that cannot be pre-processed.
- Joint analysis of citation and content for scientific document retrieval has not been fully explored. A number of research outputs have been focused on dynamic topic modelling in online social media applications without considering features that exist in scientific documents, such as citation. On the other hand, a significant amount of topic exploration works have casted their focus on the citations relationships between papers without considering much on the content [3–7].

This paper investigates new approaches to support effective scientific document retrieval by proposing a novel topic mining method that is able to discover "thematic topics" from a scientific corpus. These thematic topics have an explicit and semantic association to the research themes, and they can also contribute to effective document retrieval. The keyword set associated with each of the thematic topics needs to have the following properties:

- They should have a high co-occurrence in the documents that are relevant to the thematic topics but appear much less frequently in other documents.
- They should be not only interpretable by human researchers, but also representative and distinguishable from each other with minimal repetition and overlap between the keywords hence supporting effective retrieval of the relevant documents.

Meanwhile, topic visualisation is an important research topic under the general direction of document visualisation. Topic-based text mining methods are often coupled with interactive visualisations [1] as a promising approach to address the challenge of analysing large text corpora, allowing the users to interactively query, gain knowledge and access to the literatures through visual data exploration. Previous efforts have been concentrated on the development of numerous visualisation approaches for interactive analysis of document contents, topic structure and citation networks. The visual approaches in these works have been primarily focused on the visual analysis of citation networks, topic evolution and the creation of overall science maps. Many of them missed the important aspects of integrating the scientific environment of the publications with their dynamically changing topics.

We have designed Literature Explorer as a visual analytics suite to explore scientific literatures—see Fig. 1 for an overview. This is primarily designed to offer assistance to the researchers by helping them gather key research papers in their fields. One of the key features of the Literature Explorer is that it is capable of providing a list of key research papers that match a user-entered research topic. The search bar in the Literature Explorer allows the users to perform interactive queries about the topics of their interest and the platform then responses with a list of important papers relevant to the targeted topics. In addition, through interactive visualisation, the users can explore related topics and exam the evolution of the topics over the course of time.

From the data mining perspective, the exploration of thematic topics is based on the computation of key term occurrence in a scientific corpus. The main technical feature is a novel nonparametric key term clustering method that allows us to detect topics based on the co-occurrence of the key terms. The clustering is based on a nonparametric medoid seeking. We extend the state of the art by presenting an approach for interactive visual analysis and retrieval of scientific documents while taking into account the number of citations and their content to explore the paper relevance.

More specifically, the contributions of the paper include:

- A medoid-seeking keyword clustering method for thematic topic detection is a soft (i.e. key terms may belong to multi-topics) and nonparametric clustering without requiring the knowledge about the number of the clusters

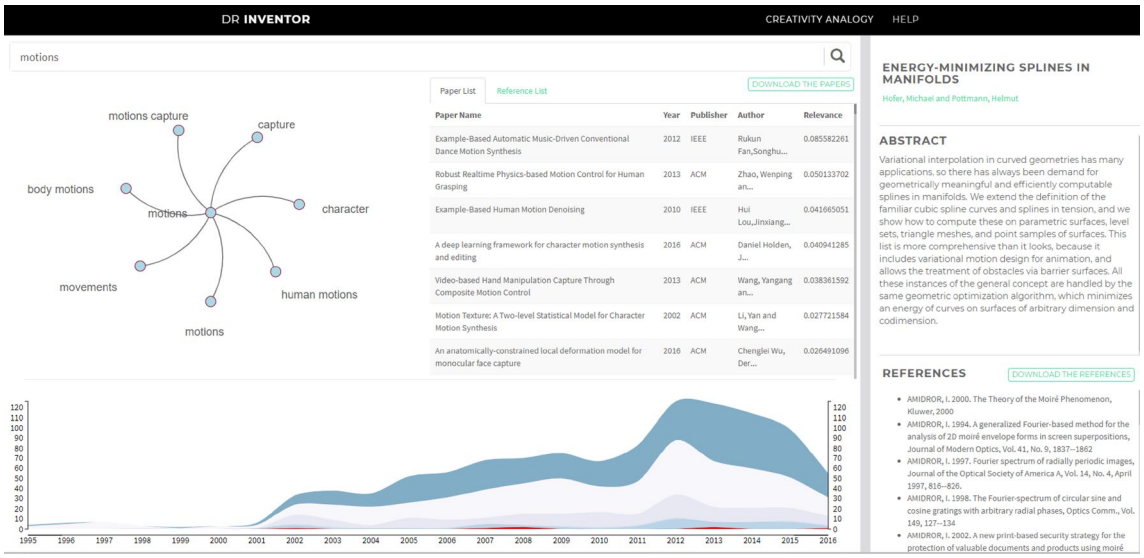

Fig. 1 Overview of the Literature Explorer: the user interface includes (1) search bar for document query by entering target thematic topic keywords; (2) graph view to display the relationships between relevant topic keywords; (3) paper list to display a list of papers that are relevant to the thematic topic; (4) theme river to show the evolutions of the relevant thematic topics; (5) paper view to display the meta-information of a selected paper

as input. The method is based on the idea of mode seeking from the density distributions in the word space. At each data point (i.e. word), the density is estimated using a density kernel. The medoids are identified according to the accumulated density by considering the density contributions from all the data points in neighbourhood. Here, we choose to use medoid as the cluster centre (i.e. modes) to avoid the needs for an explicit computation of the means. And more importantly, since each mode is associated with a keyword, it has a clear semantic meaning.

The clustering method is supported by feature representation of the key terms mined using frequent pattern (FP) mining algorithms. A word-ranking method adapted from the TF-IDF scheme is also involved to rank the importance of the keywords according to their term frequency in the entire corpus while taking into consideration the percentage of the documents that contain the words.

- A new interactive visual analytics suite allows the researchers to interactively collect papers from queried research topics. The suite consists of a set of standard visual components including trees and theme river, which are closely coupled with the underlying thematic topic detection to support effective document retrieval. The visual components are adequately integrated under the design rationale and goals. The search bar prompts relevant topic information when the user starts to enter letters and words; hence, the users can interactively search for a topic of interest. The suite then displays a paper list containing the document object identifiers (DOI) in descending order according to their relevance to the topic. The users can capture all of these papers by simply clicking on a download button. Overall, the suite offers a very convenient tool for the researchers to obtain key research papers of a target research field.

While many works in exploring scientific literatures have explored citation relations through visual data analytics, our work is focused primarily on retrieving relevant papers under a user-selected target research topic. Here from the user perspective, we stress the importance of mining the thematic topics (and hence acquiring their relevance papers) that are interpretable by human researchers. To our knowledge, there is limited work in generating scientific document collections to match a target research topic apart from those generic topic modelling methods, which suffer considerable drawbacks in terms of interpretability, usability and speed performance. We have analysed the generic methods (Sect. 2) and also compared the outcomes through evaluations (Sect. 6).

2 Related work

There are four major areas related to this work, including topic modelling from text mining, visual analytics for topic modelling and visual analytic of scientific documents. This section provides a brief summary about these areas.

2.1 Topic modelling

The topic modelling methods have become increasingly important in order to handle the ever-increasing amount of document data. Most of these methods are statistical based to mine text patterns and themes from a document corpus. The discovered topics can play a very important role for statistical analysis of document collections [8].

Latent semantic indexing (LSI) [2] was one of the earliest topic modelling methods, which applies matrix factorisation (namely singular value decomposition—SVD) on the term-document matrix to find topics based on text variances. The topics are orthogonal to each other to capture most information in the text, which makes its results useful as a representation of documents for classifications. However, the outcome of LSI allows both positive and negative weight values of the keywords in a topic, hence is not interpretable based on human standards.

In contrast, the probabilistic topic modelling methods employ nonnegative probability distributions over keywords and topics [8]. Among them, the probabilistic latent semantic indexing (pLSI) [9] and latent Dirichlet allocation (LDA) [1] are two most popular and widely used topic modelling methods, which generate topics close to human understanding by grouping co-occurring words together. In addition, hierarchical Dirichlet processes [10, 11] have been proposed to extract evolving topics from a text stream. However, topics generated by LDA may not necessarily be optimal for representing document classifications since the topics may be very close to each other to distinguish the documents between the topics. In addition, another disadvantage of these methods is their high-performance requirements.

More recently, nonnegative matrix factorisation (NMF) [12] was proposed as another topic modelling approach [13], which is based on nonnegative matrix factorisation. As all the values are nonnegative, it does not suffer from the interpretation difficulties as LSI does. Furthermore, compared to the probabilistic methods (e.g. LDA), NMF has shown its advantages in terms of running time. However, the main drawback of NMF is its inconsistency in the results when we increase the number of topics.

The topic modelling approaches are often performed together with key phrase extraction using techniques such as TextRank [14]. TextRank is a graph-based ranking algorithm that computes the importance of each vertex in the graph through voting from the vertices in neighbourhood. The voting decides the importance of each vertex, and an important vertex also casts votes with more weights. By applying this procedure to the key phrases in documents, we can identify important words for topic modelling.

There is also a significant amount of research taking place to explore topic changes over times on social media. A lot of recent efforts have focused on the analysis of topic evolution patterns in text data, including topic birth, death, splitting and merging [15]. MemeTracker [16] was developed to effectively identify phrase-based topics from millions of news articles. Some efforts have also been made recently to mine hierarchical topics and their evolving patterns in temporal datasets. For example, the evolutionary hierarchical clustering algorithm [17] generates a sequence of hierarchical clusters. However, text data from social media and from scientific domain share little commonality. The main purpose of this work is to identify publications that are related to the topics mined from the scientific literatures that are clearly understandable by the researchers.

One of key problems in topic modelling is to work out the number of the topics. In the traditional topic modelling, this is typically determined according to the size of the text corpora, either manually by human or nonparametrically by the models [18]. Potentially larger corpora lead to more hidden topics—typically tens of thousands of articles will require topics in the scale of hundreds. However, such large topic number is not user-friendly for human interpretations. Some efforts have been made to address this challenge by organising the topics in a hierarchical structure as a scalable solution to improve human interpretability. To this end, Blei et al. [1] have proposed a hierarchical topic model (hLDA) that extracts topic hierarchies from growing data to accommodate a large number of topics. However, we need to predefine the depth of the hierarchies hence the outcomes from hLDA are rather rigid. In addition, the higher level topics generated by hLDA usually consist of many stop words that are less meaningful for human users. Dou [19] proposed Topic Rose Tree, which constructs a multilevel hierarchical structure with any given number of generated topics. It leverages the scalable hierarchical structure without enforcing rigid restrictions on the topic models. However, there is no evidence of a test on scientific corpus.

2.2 Word embeddings

Word embeddings have received significant attentions in recent years. They provide high-quality vector representations that capture the semantic meanings and relations between the words. Representative embedding models include SkipGram, Continuous Bag of Words [20] and Glove [21]. Typically, the word embeddings models use a small window to capture the context words that fall within the neighbourhood of each target word.

The thematic topic detection method presented in this paper explores the similar idea, but it captures the word representations according to word co-occurrence in documents rather than in neighbourhood—see Sect. 4. This is because our interest is placed on thematic topics within documents

rather than the semantics of individual words. Also, scientific keywords that frequently appear together in the same scientific documents (but not necessarily within immediate neighbourhoods) provide important contextual information.

2.3 Visual analytics for topic modelling

Topic modelling has long been used alongside visual analytics methods for text analysis. Especially, visual analysis of evolving topics has been widely studied over years. Many of the existing methods prefer a river metaphor to convey the topic evaluation over time—for example,

TIARA [22] is one such system, and it shows the topical evolution of streaming document data. And then, TIARA adopted the visualisation in a theme river style [23] to show topical evaluation of streaming document data.

FacetAtlas [24] explored multifaceted relationships between topics in a graph layout; The work from TopicPanorama [25] visualised the links between topics from document corpora; ParallelTopics [26] showed topic evolution over time based on theme river and they also used parallel coordinates to visualise the probabilistic distribution of documents over different topics; and TextFlow [27] allowed the users to visually analyse topic merging and splitting relations by tracking their evolution over time. Xu [28] allowed interactive exploration by the users to understand the dynamics and relationships among topics. Sun et al. [29] also studied the visualisation of the relationships among topics in terms of cooperation and competition.

Often, high-performance demands prohibited real-time computation of topic models. Hence, while most of the existing system are interactive, they are not dynamic—a few exceptions include the TopicNets [30], which allowed recomputation of topic models from a dynamically changing subset of documents formed via user navigation; iVisClustering [31], which recomputed topic models on a document subset where noisy documents can be excluded; UTOPIAN [32], which offered direct steering of the topic modelling in terms of changing the keyword weights of a topic, splitting and merging topics, and creating a new topic based on a seed keyword or document. In comparison with the text data from social media, scientific research topics are much more steady and new research topics only appear at a much slower pace. Hence, while the Literature Explorer does possess the capacity of automatically detecting the topics without prefixing the numbers, our primary focus was to answer the key research question how to acquire a list of the most relevant scientific publications according to the user query on a selected research topic, rather than to dynamically update the topics.

The Literature Explorer frontend on client side is designed as a web application which is supported by AngularJS to gain the operational convenience and speed. The data operations are designed to be balanced on both client side and server side, so that the highly dynamic interactions for the enquiring and visualising can be achieved.

2.4 Visual analytics of scientific literatures

In the past, many methods have been developed to utilise the massive amount of available scientific literatures, including those measuring the prolificacy of authors [33] and detecting research trends. Many papers have applied data mining techniques to analyse the literatures [34] and summarise disciplines [35]. A lot of focuses have been paid towards the bibliographic data, leading to either science landscapes [36], or the exploration of citation networks, or the interrelationships of authors. CiteSpace II [37] is an approach that is based on citation network analysis. IN-SPIRE by Wong et al. [38] is a text analytics tool to identify research topics over time; Heimerl et al. [39] presented an interactive visual approach for scientific literature classification; and the PaperLens system [40] is another system with similar features. It also shows the most often cited authors and papers every year. CiteRivers [41] focused on citation by presenting a new representation of citations based on community structure and the underlying topics. Examples of the recently published work on visual approaches for scientific literature browsing and search based on topic exploration include the Action Science Explorer [42], which was designed to structure and analyse which a collection of scientific documents for literature overview, and Beck et al. [43], which supported paper search and key paper identification through the structure of citation network; ThoughFlow [44], which visualised literature collections using topic models to bridge the information gap between activities for research idea generation; Cite2Vec [45], which presented a citation-driven document exploration through word embeddings. However, very few work aims to facilitate joint analysis of contents, topics and citations, leaving a crucial analysis gap here.

In contrast, Literature Explorer was designed to support interactive literature exploration via joint topic analysis through content (namely occurrence of key words) and citations. The underlying technique, namely the nonparametric topic detection, yields outcomes (i.e. topics represented by a list of keyword) that strike a balance between the uniqueness, completeness and interpretability of the topics. The generated topics from our method are distinguishable to each other, interpretable by humans, and they collectively cover the whole spectrum of the selected topics (similar to the property achieved by LSI). These properties allow the researchers to clearly identify their topics of interest, and subsequently support the identification and retrieval of the most relevant papers. The evaluation indicates a better performance of our method as compared to the literature search based on the conventional topic modelling methods.

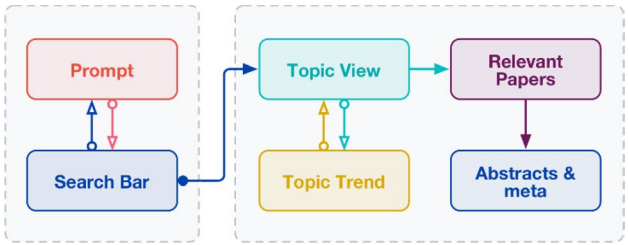

Fig. 2 System operational flows

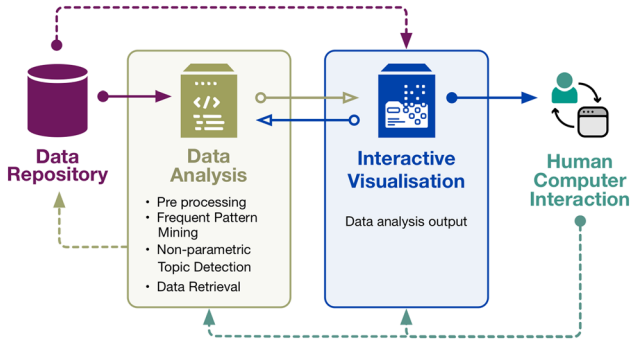

Fig. 3 The system overview of Literature Explorer

3 Overview

3.1 Use scenario

Literature Explorer is designed to help academic researchers gather key research papers according to interactively selected research topics. The user interface is given in Fig. 1. Figure 2 depicts the operation flow. The entry point is the search bar. The main use scenario is described as follows:

The user uses the search bar to enter thematic topic keywords. S/he can start to type in a few words; then, the related topics are prompted in a drop down list. The inputs can be selected from the prompted list. The target keyword is then highlighted in red colour, and the user interface is automatically updated according to the new selection. Then, the user is able to review:

- Keywords related to the target thematic keyword presented through the interactive graph view,
- Papers related to the target keyword presented in the paper list. The list can be sorted according to the titles: year, author and relevance in an ascending or descending order. Each paper can be selected to show its abstract and references in the Paper View. In addition, there is a download button to download the paper list.
- The evolution of these relevant keywords presented through the theme river.

At any time, the user can restart the search by entering new keywords in the search bar. More details about the user interface and interaction are provided in Sect. 5.

3.2 Data sources

Scientific publications are often well-structured, typically including key information such as title, abstract, references and so on. Without loss of generality, the outcomes of the work are demonstrated in the domain of computer graphics. The corpus involved in this work includes thousands (i.e. 3589) of research publications from ACM [46] and IEEE [47] in the computer graphics domain in the last 22 years, which represent a major resource of scientific literature used by the computer graphics researchers to advance the research in the area.

3.3 System components

We design Literature Explorer as a visual analytics platform to explore research literatures. The system is composed of four modules, namely the data repository, data analysis, visualisation and interaction module—see Fig. 3. The data repository accommodates the meta-information about the scientific publications, and it also saves the results from the data analysis. The data analysis module includes pre-processing the raw data, frequent pattern mining, nonparametric topic detection and relevant paper retrieval. The visualisation presents the output from the data analysis module. The user interaction allows for interactive topic queries, topic selection, paper selection and paper download, hence allowing the users to explore the target topics and collect papers that are highly relevant to the target topics.

The system is designed for flexible scalability. Between the four modules, APIs are used for communication and data exchange. In between the server and clients, standard HTTP and RESTful APIs are applied for the data queries. The data analysis module is constructed in a sequential mode, and the data is processed in a pipeline by the system including text pre-processing, frequency pattern mining, topic detection and visualisation. The data is obtained from the data repository. The intermediate results of each stage are stored in the database. The final topics are stored in the database for query from the client side.

The data pre-processed involves a number of steps, including format conversion, cleaning (e.g. removing incorrect words, etc.) and information extraction (e.g. metadata, reference, DOIs). The cleaned texts are stored in the database for further processing. The references are parsed for further use. The word occurrence is counted, and sentences and paragraphs are split.

Frequent pattern mining is performed based on the FP-growth method [43]. This is to identify the co-occurrence of

the keywords for topic detection. (More details about the topic detection are presented in Sect. 4.). Practically, this involves the use of the scalable Machine Learning Library from Apache Spark,[1] which uses in memory and multiple thread computations. This computation is needed only once for the preparation of a single corpus.

The data repository module receives raw data from different sources. The data source is described in Sect. 3.2. Further documents can be added into the system to support the query of more scientific papers. The thematic topic modelling only needs to be processed offline once based on the collection of sufficient documents in the corpus. Newly added documents can be matched to the detected thematic topics for the query.

The visualisation is implemented in D3.js, and the control and interactive parts are implemented in AngularJS.

4 Nonparametric thematic topic detection and paper retrieval

The paper retrieval in the system is based on the detection of thematic topics from the corpus. This is a nonparametric approach which does not need to pre-define the topic number. The aim is to detect topics that are meaningful according to the conventions in the research communities and also distinguishable to each other. The nonparametric topic detection is based on a new keyword ranking scheme (Sect. 4.1) and the feature representation as described in Sect. 4.2. The topic detection is essentially carried out based on medoid seeking (Sect. 4.3).

4.1 Word ranking

Given a collection of scientific corpus from a specific domain, we utilise an adapted TF-IDF scheme to extract domain words and rank their importance according to their total frequency in the corpus, taking into consideration the percentage of the documents that contain the words in the entire corpus. This scheme is based on two observations:

- Important words tend to have a high occurrence
- Important words tend to NOT appear in all the documents from a domain—this is because there are a range of research topics in the domain and often papers were published to target a set of specific research topics only.

Based on the above observations, we create a ranking score matrix $\pi(w, d)$, which represents the ranking of each word w in document d, and $\pi(w)$, which is the total score of word w as follows:

$$\pi(w, d) = (\mathrm{fq}(d, w)) \times \mathrm{idf}(w) \times \mathrm{ca}(w) \tag{1}$$

$$\pi(w) = \sum_{|D|} \pi(w, d) \tag{2}$$

where $\mathrm{fq}(d, w)$ is the occurrence frequency of a word w in document d. Further, $\mathrm{idf}(w) = n(w)/N$ is the inverse document frequency of w, where $n(w)$ is the number of documents that contains word w and N is the total number of the documents in the corpus; $\mathrm{ca}(w)$ is used to remove the influence from the words with very low frequency, and its parameter T is set empirically.

$$\mathrm{ca}(w) = \begin{cases} 1 & \text{if } n(w) \geq T \\ \frac{n(w)}{T} & \text{other wise } n(w) < T \end{cases} \tag{3}$$

The word ranking is used in the topic detection in Sect. 4.3 as an importance indicator of the words.

4.2 Feature representation

We treat the key terms extracted from the corpus as random variables and subsequently model the relationship between them using an undirected graph $G=(V, E)$, where V is the set of nodes, each of which represents a key term, and E is the set of edges between the key words. The word relationship is modelled via their co-occurrence. Two nodes are linked if and only if the words that they represent have co-occurrence in the documents within the corpus. Here, we define the strength of the link as follows:

Definition 1 *Word Link Strength*: The strength of the link between two words $\mathrm{w_s}$ and $\mathrm{w_t}$ is defined as:

$$\sigma(w_s, w_t) = |\{d_i | w_s \in D_i, w_t \in d_i, d_i \in C_p\}| \tag{4}$$

where the symbol $|\cdot|$ denote the number of elements in a set and d_i is a document in the corpus C_p. In essence, this is the total number of the documents that contain both words w_s and w_t

For a variable $s \in V$, let $\Gamma(s)$ denote the neighbours of s. According to the local Markov property, for any two variables $s, t \in V$, the variable s is independent of t conditioned on its neighbours $\Gamma(s)$, suggesting that the word represented by s can be characterised by $\Gamma(s)$, as the neighbours contain all of the information necessary to decide its value. Note such a co-occurrence property between words is also exploited in other applications such as word embedding in which the word semantics are explored according to their co-occurrence within the word neighbourhood. According to this property, each term associated with s can be represented by a vector $\mathbf{X}_s$ using $\Gamma(s)$ as the features components and $\sigma(w_s, w_t)$ as the component weight.

[1] https://spark.apache.org/docs/1.1.1/mllib-guide.html.

Hence, the distance between two words w_s and w_t is defined as follows:

$$d(w_s, w_t) = d(\mathbf{X}_s, \mathbf{X}_t) \times p(w_s, w_t) \tag{5}$$

where the first term $d(\mathbf{X}_s, \mathbf{X}_t)$ in the equation stands for the distance function between the two vectors, which is calculated using one minus the cosine similarity (normalised). The second term in the equation stands for the joint probability of w_s, w_t, which can be calculated as

$$p(w_s, w_t) = \frac{\sigma(w_s, w_t)}{N} \tag{6}$$

where $\sigma(w_s, w_t)$ is calculated using Eq. (4), and N is the total number of the documents in the corpus. Hence, we measure the word distance based on not only their similarity, but also how likely they co-occur in the same documents.

Practically, we use the FP-growth method presented by Han [43] to capture the co-occurrence of words. It employs a divide-and-conquer strategy to mine the frequent patterns without involving candidate generations.

4.3 Nonparametric topic detection

The proposed medoid-seeking keyword clustering method is a soft (namely a word may belong to multiple clusters) and nonparametric clustering without requiring prior knowledge about the number of the clusters. The method is designed to identify cluster centres as well as to infer the probability of the words belonging to the centres. It is based on the idea of mode seeking from the density distributions of the data. At each data point, the density is estimated using a density kernel—here we use Gaussian function without loss of generality. Medoids [48] are identified according to the accumulated density by considering the density contributions from all the data points in neighbourhoods. These medoids are treated as the cluster centres. An initial computation of the densities around the cluster centres is performed, which is further updated iteratively by the term-to-cluster probability and the cluster density in an alternated manner.

A medoid is defined in the word space as the most centrally located point among a set of sample points. It has the minimum distance with all the other samples. We use medoids because each of the medoid associated with the underlying mode is discovered from an individual keywords, leading to an explicit semantic meaning to the medoid. Also, they can be considered as a good representative of their neighbours while do not require computation of the mean of the sample points.

4.3.1 Density functions and medoid

Definition 2 *Word Density*: We define the density at word w_j in the form of an accumulation of mixed contributions from the density distribution of other words as:

$$f(w_j) = \sum_{i=0}^{n} p(w_i) \times f_i(d(w_i, w_j)) \tag{7}$$

where n is the number of all the keywords, $p(w_i)$ can be obtained using $\pi(w_i)$, namely the importance score from Eq. (2), as follows:

$$p(w_i) = \frac{\pi(w_i)}{\sum_{j=1}^{n} \pi(w_j)}. \tag{8}$$

And $f_i(d(w_i, w_j))$ is the probability density function based on the distance value d between two words w_i, w_j. In other words, it represents the contribution of the density function of w_i towards w_j based on the distance between these two words. We use Gaussian kernels to estimate the densities. The size of the Gaussian kernel is estimated in an adaptive manner according to the data histograms. This is further moderated by the probably of w_i, which is $p(w_i)$. Hence, a more important word has more influence on its neighbours.

Once the density at each w_j is calculated, we can identify an initial set of the medoids by finding the local maximums, leading to a set of medoids C_k as follows:

$$C_k = \{w_k | f(w_k) > f(w_j), \quad \forall j \in \Gamma(w_k)\}, k = 1 \ldots K \tag{9}$$

This nonparametric procedure allows us to find the medoids through local maximum, namely those nodes with the highest density values among their neighbours. The number of the clusters depends on the number of the local maximums.

4.3.2 Clustering

Upon the identification of the medoids, which are treated as cluster centres, we compute the word clustering through iterations. This is a soft clustering in the sense that we compute the probability of a word w_i in each cluster C_k. The iteration starts from the following initialisation, which computes the probability purely based on the distance between w_i and C_k:

$$p(C_k | w_i) = \frac{1 - d(C_k, w_i)}{\sum_{k=1}^{K} (1 - d(C_k, w_i))} \tag{10}$$

$$d(C_k, w_i) = d(w_{C_k}, w_i)$$

where w_{C_k} stands for the centre word (i.e. medoid) for C_k and K is the total number of the clusters according to the local maximums. Hence, initially we assign the probability of a word to different medoids purely based on the normalised distance. This is to be iteratively updated below.

With the initial probability computed from Eq. (10), we compute the cluster density for each cluster k at w_j as the sum of the density contribution from all the other words at w_j:

$$g_k(w_j) = \sum_{i=0}^{n} p(C_k|w_i) \times f_i(d(w_i, w_j)) \times \text{rank}(C_k) \tag{11}$$

where n is the number of the words, $\text{rank}(C_k)$ denotes the importance ranking of C_k, which is initialised as 1 for all the clusters, and this is updated iteratively via Eq. (13)—see below. Hence, the density contribution from cluster k to word w_j is contributed by all the words according to their distance with further moderations based on the cluster ranking C_k and the probability that measures how much these words belong to C_k.

With $g_k(w_j)$, we compute $p(w_i|C_k)$ using normalised densities of all the n words at cluster k. More specifically, this is done as follows:

$$p(w_i|C_k) = \frac{g_k(w_i)}{\sum_{i=1}^{n} g_k(w_i)} \tag{12}$$

Further, we update the importance ranking of each cluster k according to the importance of the words and how much they belong to the cluster:

$$\text{rank}(C_k) = \sum_{i=0}^{n} p(C_k|w_i) \times \pi(w_i) \tag{13}$$

Similar to Eq. (12), here we use normalised densities of all the clusters to update $p(C_k|w_i)$ as follows:

$$p(C_k|w_i) = \frac{g_k(w_i)}{\sum_{k=1}^{K} g_k(w_i)} \tag{14}$$

The steps between Eqs. (11) and (14) are repeated until there is no further change in the results. This provides us with a collection of words w_i ranked according to the conditional probability $p(w_i|C_k)$ for each cluster C_k, which collectively represent topic k.

Algorithm 1: Non-parametric thematic topic detection

input : A scientific document corpus from a subject domain
output: A collection of words w_i ranked according to the conditional probability $p(w_i|C_k)$ for each cluster C_k

```
Function nptTopicDetection(data)
    // Compute the accumulated density for each
       word w_j using Equation (7)
    foreach word in document do
        accDensity ← accumDensity(word);
    end
    forall k in C_k do
        initRank(C_k, 1)
    end
    initCondProbability ← calculateProbability(C_k, w_i);
    foreach cluster(k) do
        // Compute the cluster density for each
           cluster k at w_j using Equation (11)
        clusterDensity ←
          calculateClusterDensity(k, w_j);
        // Compute the conditional probability
           p(w_i|C_k) using Equation (12)
        condProbablity1 ← calculateProbability(w_i, C_k);
        // Compute the ranking of each cluster k
           using Equation (13)
        clusterRank ← calculateRanking(k);
        // Update the conditional probability
           p(C_k|w_i) using Equation (14)
        condProbablity2 ← updateProbablity(C_k, w_i);
        // Repeat the loop until there is no
           change in p(w_i|C_k)
        if no change in condProbablity1 then
            break;
        else
            continue;
        end
    end
    return accDensity;
end
```

In addition, the similarity between the topics is measured based on their associated keywords. Each topic can be represented by a vector, in which each dimension is weighted according to the conditional probability $p(w_i|C_k)$. Their similarity is measured based on the cosine similarity.

Once the topics are discovered, the relevance of a paper to a topic can be calculated according to how much the words in the paper belong to the topic and the importance of these words:

$$p(C_k|d) = \frac{\sum_{i=0}^{n} n(w_i, d) \times p(C_k|w_i) \times \pi(w_i)}{\sum_{k=1}^{K} \sum_{i=0}^{n} n(w_i, d) \times p(C_k|w_i) \times \pi(w_i)} \tag{15}$$

where d is a paper and the equation computes the relevance of the paper to the topic C_k, $n(w_i, d)$ is the occurrence of word w_i, in d.

Further, from a given topic, we calculate the relevant papers based on the importance of the words in the topic and the percentage of the words:

$$p(d|C_k) \sim \sum_{i=0}^{n} n(w_i, d)/n(d) \times p(w_i|C_k) \tag{16}$$

This relevance score forms the basis of the paper ranking for the given topic. Hence, a list of ranked papers (so called paper list) can be discovered according to their relevance scores to the topic.

In addition, based on this ranking calculation, we also compute another ranking score based on the citation numbers. The basic idea is to give a paper higher ranking scores if it is cited by more papers from the paper list. The calculation is straightforward—for each paper in the paper list, we exam its citations and cast a vote to each paper in the reference list. Once we complete the processing of all the papers in the list, we can have a list of papers (so called Reference List) ranked according to the votes they have received.

Notably, the citation information provides a very important metric to the paper ranking in addition to the text. Practically, we have also applied the mining to the other parts of the scientific documents, but the results are very similar. Hence, empirically we found that using abstract and citation provides a good balance between the outcome and the computational demands.

5 Visualisation in Literature Explorer

We use visualisation to support the users to leverage the outcomes of the new thematic topic detection method described in Sect. 4. The platform offers simple and clear information representation to facilitate user understanding and interaction by exploiting visual elements that are closely coupled with the data mining method to fulfil the design rationale and goals.

5.1 Design rationale and goals

Our design goal is to support the academic researchers in literature review by automatically gathering the relevant papers in an interactively selected research topic, with particular focuses on two goals, namely the *simplicity* and *clarity*. The former is related to the easy access to a collection of key papers for a target research topic in the area without requiring too much manual effort, while the latter refers to the presentation of the topics and papers with clear layout and easy-to-read information.

As Literature Explorer is designed for scientific research, the targeted end-users are academic researchers. They are facing a great challenge arising from the fast growing volume of scientific literatures. Hence, there is a significant demand in having access to relevant research information and resource in a timely and efficient manner. To reflect this, we employ strategies to ensure the design of a simple and intuitive user interface.

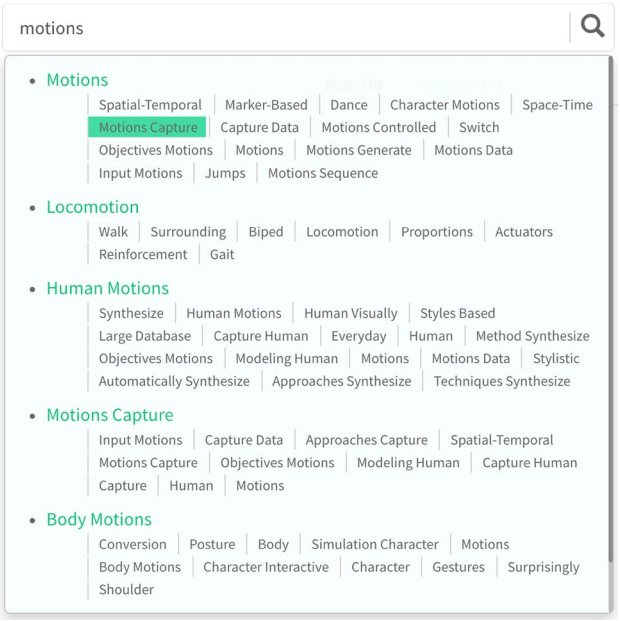

Fig. 4 An illustration of search bar

5.1.1 Simplicity

Paper collection (G1) The users need to collect important papers in a research topic of their interest. They want to achieve this through interactive queries using relevant keywords.

5.1.2 Clarity

Relevant topics (G2) The users need to gain an overview about the relations between the target keyword and other relevant keywords.

Thematic topic evolution (G3) The users need to see the trend of the research topics by viewing the numbers of relevant publications over time.

5.2 User interface and interaction

5.2.1 Search bar

The search bar allows the users to interactively explore research topics of their interest. It also includes

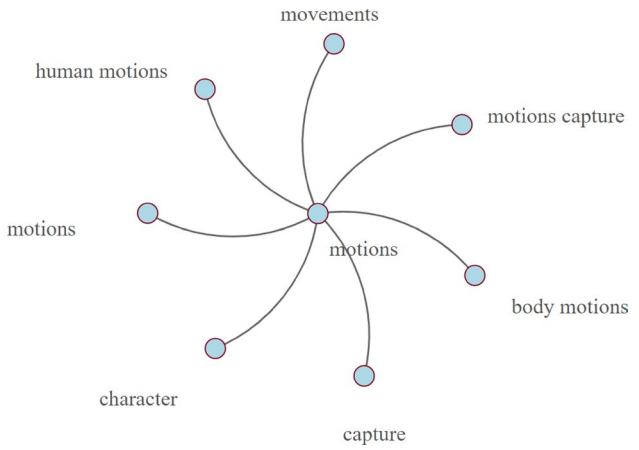

Fig. 5 An illustration of graph view

autosuggestion to help the text entering—the system automatically suggests a list of relevant topic keywords based on the user input for interactive selection. An illustration of the search bar is given in Fig. 4.

5.2.2 Graph view

The graph view displays the relationship between different topic words using a simple graph. The users can interactively select a topic from the graph to further explore its relevant research papers. An illustration of the graph view is given in Fig. 5.

5.2.3 Paper list and reference list

The paper list offers a list of papers that are relevant to the selected research topics. These papers are obtained via Eq. (16). The users can choose the order of the paper list according to their relevance score and the year of publication, etc. Similarly, the reference list offers papers that are discovered and ranked according to their citations by the paper list. An illustration of the paper and reference list is given in Fig. 6. The users can also click the download button to capture the paper list in a separated file. Each paper in the list includes its DOI information, which allows the users to access its full paper information very easily—see Fig. 7.

5.2.4 Theme river

The theme river displays the evolution of a set of related research topics along a timeline. The topics are identical to those displayed in the graph view. The users are also able to make selection of the topics from the theme river. An illustration of the theme river is given in Fig. 8.

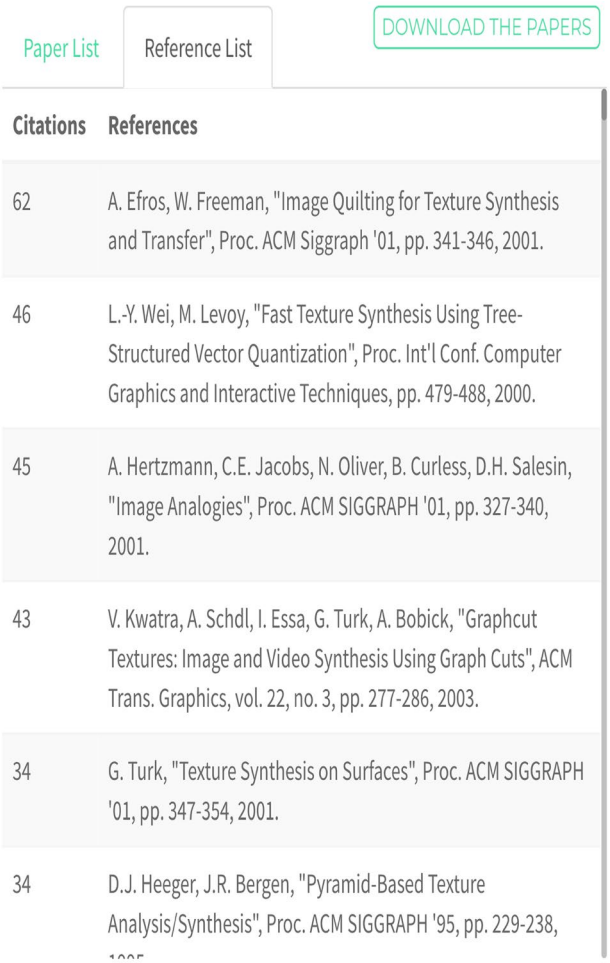

Fig. 6 An illustration of paper list (left) and reference list (right)

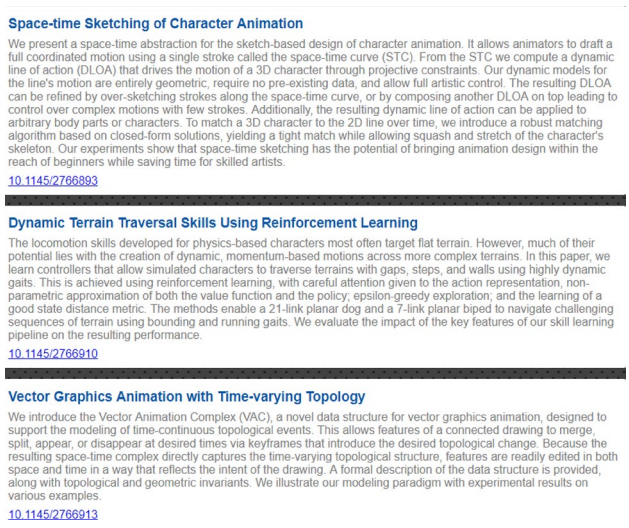

Fig. 7 An example of downloaded paper list

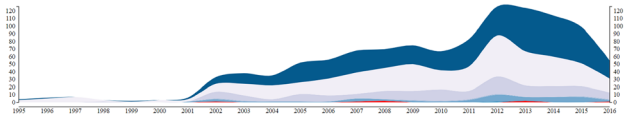

Fig. 8 An illustration of theme river

5.2.5 Paper view

If the user selects a paper from the list, the paper view displays the title, authors, abstract and the reference of the selected paper. Figure 9 shows an illustration of the paper view.

6 Evaluation

Literature Explorer is evaluated through the combination of involving human participants and carrying out objective measurements. Without loss of generality, we collect a corpus of computer graphics documents for the evaluation. The documents and related metadata are stored in a database. The documents are mainly composed of the conference papers from SIGGRAPH [46] and from IEEE Transactions on Visualisation and Computer Graphics [47]. Totally, the corpus contains 3589 full paper documents.

The evaluation was conducted to ratify how Literature Explorer supports individual researchers in literature review in the domain of computer graphics by designing and completing tasks. In total, eight participants were involved in the evaluation. The evaluation team was consisted of either post doctoral researchers or senior PhD students in computer graphics. They were instructed to use the system, assess the system in research topics that they already knew by comparing the system output (i.e. the paper collection) against their expert knowledge. All of them completed the designed tasks described in Sect. 6.1 and the usability questionnaires. The evaluation scores provide the assessment of the proposed method and platform. The evaluation design for the visualisation was based on the assessment against the design goals of visualisation in Simplicity and Clarity as discussed in Sect. 5.1—see Table 1 for more details. All the results are given in Sect. 6.2.

To measure the matching (alignment) between the system retrieved literatures and the expert knowledge, we define the "Precision Rate (PR)" as the rate of the number of the "corrected papers" versus the total number of the papers retrieved by the system.

Similarly, to measure the coverage of the retrieved literatures from the system in a given topic area, we define the "Recall Rate (RR)" as the percentage of the important papers in the topic area that are discovered by the system. Practically,

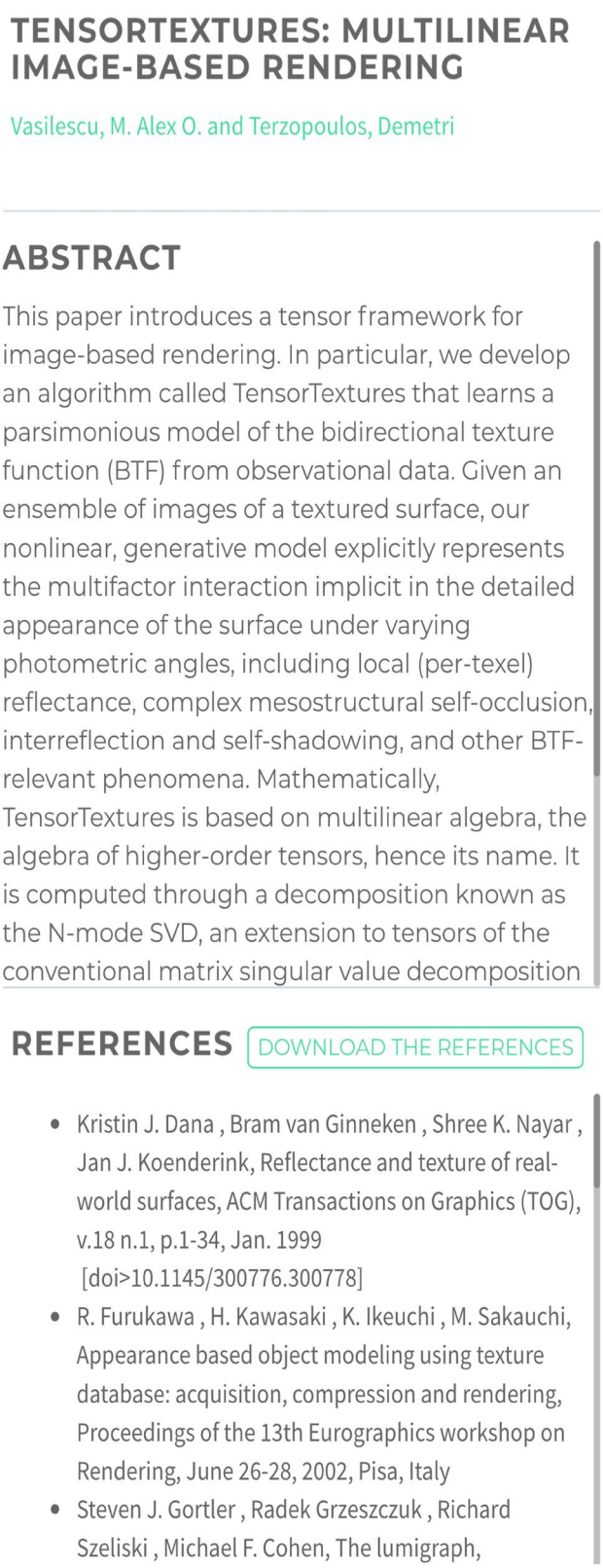

Fig. 9 An illustration of paper view

Table 1 Evaluation for the visualisation

	Requirement	Visual components	Evaluation methods
Simplicity	G1: Paper collection	Paper list Reference list Paper view	Precision rate Recall rate
Clarity	G2: Relevant topics	Graph view	Usability questionnaires
	G3: Topic evolutions	Theme river	Usability questionnaires

we asked the evaluation experts to provide a list of "important papers" for the target research topics for the evaluation.

In addition to the evaluations by the experts, we also carried out objective measurements by comparing the outcomes from Literature Explorer against three other topic modelling methods, including LDA, NMF and TextRank. The results are presented in Sect. 6.2.

6.1 Tasks

To assess the PR and RR in the document retrieval, we design the following tasks for the participants:

- Use search bar to enter a thematic topic word at the choice of the participant.
- Prepare "My List", which is a list of important papers of the selected topic according to the knowledge of the participants, which they may find from their own literature reviews.
- Make sure the target topic word is selected—the selected word is highlighted in red colour
- Evaluate the "paper list"
- Indicate how many papers on the list are relevant to the selected topic (for PR)
- Indicate how many papers in "My List" are in the "paper list" (for RR)
- Repeat the same practice as above for the "Reference List"

7 Results

7.1 PR and RR

We define PR as the percentage of the papers from the system output that are deemed to be relevant by the expert participants, and the RR as the percentage of the papers from "My List" that are recovered by the system (namely appeared in the paper and reference list in the system). Table 2 shows the results of PR and RR.

Eight topics were evaluated by the participants. By topic searching, Literature Explorer provides a relevant paper list and a Reference List which are associated with the searching topics with average PR of 76.1% and 89.7%, respectively, and average RR 88.1% and 75.0%, respectively. For the paper list and Reference list in RR, they may have overlap for the statistics. The total is the union of the statistics from the Paper and Reference list, which is greater than or equal to the better one. The average of the Total result is 88.3%.

Table 2 Evaluation (precision rate and recall rate)

Evaluated topics	Precision rate		Recall rate		
	Paper (%)	Ref. (%)	Paper (%)	Ref. (%)	Total (%)
Facial animation	100.0	93.3	100.0	85.7	100.0
Hair modelling	62.5	85.7	72.2	77.8	83.3
Mesh deformation	53.3	72.7	80.0	86.7	86.7
Human modelling	75.0	77.8	71.4	42.9	85.7
Skin simulation	75.0	93.8	62.5	54.5	72.7
Tree modelling	87.0	100.0	100.0	100.0	100.0
Physical simulation	100.0	100.0	100.0	50.0	100.0
Mesh simplification	50.0	65.0	75.3	86.7	93.3
Average	76.1	89.7	88.1	75.0	88.3

In general, PR has better results than RR—as the system only handles SIGGRAPH and IEEE TVCG papers at the moment.

The Reference List is always better than the paper list. This is because the Reference List is inferred based on the common citation between papers in the same topic; it does provide a good measure for retrieving papers in a target topic.

The searching result is 88.3% with the combination both paper list and reference list, which means most of the selected documents can be queried from the Literature Explorer platform given their topics. For three topics: "Skin", "Tree modelling" and "Mesh simplification", the recall rates are low. The reason is that these papers are not covered by the current dataset.

Table 3 Evaluation of the platform usability

	1	2	3	4	5	Avg.
The software performs the intended tasks				6	7	4.5
The functionalities involved in the system are sufficient				9	4	4.3
The system is able to give expected results			2	3	8	4.5
The system interacts quickly		1	4	4	4	3.8
I can comprehend and learn to use the system easily			2	9	2	4.0
The interface looks good & provides all required information			4	9		3.7
Usage of the system is intuitive			3	9	1	3.8
The software is capable of handling errors		1	6	6		3.4

Agree levels: 1—strongly disagree, 2—disagree, 3—neither agree nor disagree, 4—agree, 5—strong agree

7.1.1 Usability

For the usability, Table 3 provides the distribution of the answers from the participants to each of the questions, as well as the average score for each usability questionnaire (i.e. the last column). The scores are set from 1 to 5, the higher is better. The total average score is 4.0. Four scores are more than 4.0, and other four are more than 3.4.

While most of the participants gave positive feedbacks on the usability, clearly there is space to improve the performance and the outcomes.

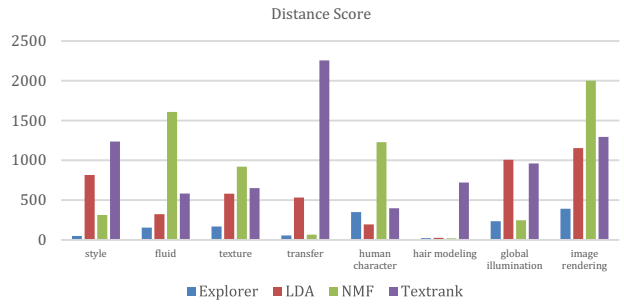

Fig. 10 Distance scores of different topic modelling techniques

7.1.2 Comparison with LDA, NMF and TextRank

The outcomes from Literature Explorer have been compared with three other topic modelling algorithms: LDA, NMF and TextRank. The Literature Explorer has the best performance—see more details below.

From the collected corpus, the nonparametric thematic topic modelling in Literature Explorer has identified 1379 topics. We applied the same number to the other topic modelling algorithms for comparison.

In addition, for LDA, we set the following hyperparameters: $\alpha = 50/K$ for the topic document generation, where $K = 1379$ (the topic number); and $\beta = 0.01$ which is an empirical value for topic word generation. For NMF, we built an 8063×3589 matrix, where 3589 is the document number, and 8063 is the keyword number, to indicate the occurrence of the words in each document. The rank is set as 1379 (i.e. the number of topics).

In TextRank, we use sentences rather than documents as the basic element to establish the topics. A similarity matrix is established to describe the relations between sentences, and a graph is constructed based on which the score for each of the vertices is calculated. Hence, each score is given to a sentence. The top 1379 scored vertices are used as the retrieved topics. The word set (with removed stop words) in the sentence are used as topic members.

Eight important research topics in computer graphic are selected for the evaluation by comparing the outcomes from this paper against those from other topic modelling methods, namely LDA, NMF and TextRank. More specifically, we compare the list of the papers retrieved from Literature Explorer, and those from LDA, NMF and TextRank.

These lists are compared based on their total distance (i.e. distance score) to each of the eight topic words (see Eq. 5), which indicates the relevance of the retrieved papers to the topics under query. A smaller distance score gives higher ranking (relevance). This is based on the observation that a better retrieval algorithm should come up with a list of papers with closer distance to the thematic words. Note our focus here is to compare the outcome of paper retrieval based on the topic modelling rather than the topics generated by the topic modelling methods.

As shown in Fig. 10, the papers from Literature Explorer generally have better results (i.e. smaller scores) compared to LDA, NMF and TextRank. Literature Explorer has similar performance as NMF in topics "transfer" and "global illumination" but is much better than LDA and TextRank. For "hair modelling", Literature Explorer, LDA and NMF, three methods have similar good results, all of them are much better than TextRank. Only in the "human character" topic, the result from Literature Explorer is worse than LDA but still higher than NMF and TextRank. NMF has very bad performance in topic "image rendering", because it cannot find any relevant documents. Overall, TextRank has the worst performance compared to other methods.

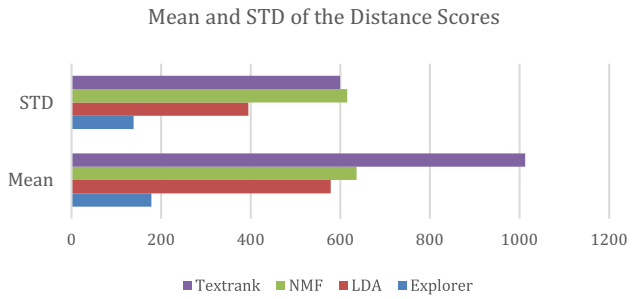

Fig. 11 Average and standard deviation of the distance scores

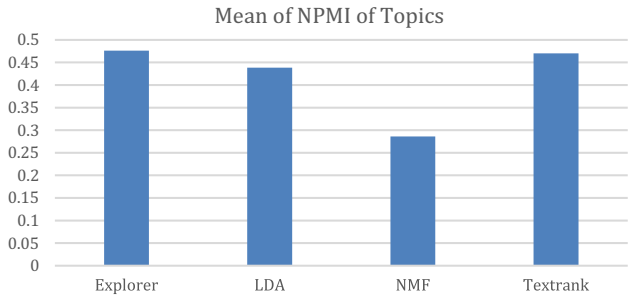

Fig. 12 The average NPMI of the topics from Literature Explorer, LDA, NMF and TextRank

Also for each topic, we average the score of the paper lists from Literature Explorer, LDA, NMF and TextRank. As illustrated in Fig. 11, our method outperforms the other three methods in general. The average ranking is 178.58 comparing to 578.6, 615.41 and 1012.5 for LDA, NMF and TextRank, respectively. Also, Literature Explorer has more stable performance according to its standard deviation 138.43, which is much lower than 394.69, 636.41 and 599.5 from LDA, NMF and TextRank, respectively. NMF and TextRank have similar results compared to Literature Explorer and LDA.

In addition, we have also compared the associations of the keywords in the topics by using pointwise mutual information (PMI) [49, 50]. The PMI between two random variable x and y is calculated as follows:

$$\text{pmi}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}. \tag{17}$$

where $p(x)$ and $p(y)$ are the probability of random variable x, and y, respectively, and $p(x, y)$ is their joint probability.

For comparison, we use the normalised PMI as follows:

$$n\text{pmi}(x, y) = \frac{\text{pmi}(x, y)}{h(x, y)} \tag{18}$$

where $h(x, y) = -\log p(x, y)$ is the definition of self-information.

We calculate the average of the NMPIs of all the word pairs in the topics. A higher value means a stronger association between the words in the topic, which implies better results. Among the eight research topics evaluated Literature Explorer has the highest average NPMI score (Literature Explorer: 0.4762, LDA: 0.4384, NMF: 0.2862, TextRank: 0.4702)—see the plot in Fig. 12.

7.1.3 Time and complexity

The detection of the frequent patterns uses the Spark Library from Apache. A large FP-tree is setup for the scanning. The minimum support number is set as 10. To process the corpus, the computation took 23 hours on Windows 10 with Intel i5 processor 2.4 GHz, 4 Core CPU and 32 GB memory configuration. But this only needs to be done offline once for the entire corpus.

The complexity of the thematic topic detection algorithm (namely Algorithm 1) is $O(T \times n \times k)$, where n is the keyword number, k is the number of topics, and T is the iteration times. This needs to be performed offline only once to detect all the topics, and the running time for our corpus is 160 min. Once the topics are discovered, the system can retrieve relevant papers within seconds. This process is on average 1.4 s for topic input, server query and returning the relevant papers for the visualisation in the web page.

8 Conclusion

Literature Explorer is designed to support academic researchers in literature review. It features a nonparametric topic detection method to detect thematic topics from a collection of scientific corpus. The detected topics have explicit associations to the research themes that are commonly used by human researchers. In addition, these topics contribute to the effective retrieval of relevant research papers that match the research themes. It also includes a new visual analytics suite that consists of a set of visual components that are closely coupled with the underlying thematic topic detection to support effective document retrieval and are adequately integrated under the design rationale and goals. The evaluation results show the comparison between the proposed method with other topic modelling approaches including LDA, NMF and TextRank. The newly proposed method can outperform the existing approaches in most of the evaluated cases. In addition, expert evaluation by the participants (researchers) has also confirmed a good usability of the system.

The future work will be focused on the extension of the corpus beyond the scientific field of computer graphics, hence increasing not only the size but also the variety of

the content in the corpus. This would allow us to further evaluate the method under a more general scientific setting, and the platform can also serve as a tool to support literature exploration by researchers from wider research communities.

Acknowledgements This research is supported by the European Commission with project Dr Inventor (No 611383), MyHealthAvatar (No 60929), and by the UK Engineering and Physical Sciences Research Council with project MyLifeHub (EP/L023830/1).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J Am. Soc. Inf. Sci. **41**, 391–407 (1990)
3. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. Presented at the Proceedings of the First Workshop on Social Media Analytics, Washington D.C., District of Columbia (2010)
4. She, J., Chen, L.,: TOMOHA: TOpic model-based HAshtag recommendation on Twitter. Presented at the Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, (2014)
5. Spasojevic, N., Yan, J., Rao, A., Bhattacharyya, P.: LASTA: large scale topic assignment on multiple social networks. Presented at the Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, New York, USA (2014)
6. Vosecky, J., Jiang, D., Leung, K.W.-T., Ng, W.: Dynamic multifaceted topic discovery in Twitter. Presented at the Proceedings of the 22nd ACM international conference on Conference on Information & Knowledge Management, San Francisco, California, USA (2013)
7. Lin, W., Pang, X., Wan, B., Li, H.: MR-LDA: an efficient topic model for classification of short text in big social data. Int. J. Grid High Perform. Comput. **8**, 100–113 (2016)
8. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**, 77–84 (2012)
9. Hofmann, T.: Probabilistic latent semantic indexing. Presented at the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, USA (1999)
10. Zhang, J., Song, Y., Chen, G., Zhang, C.: On-line evolutionary exponential family mixture. Presented at the Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA (2009)
11. Zhang, J., Song, Y., Zhang, C., Liu, S.: Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. Presented at the Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
12. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788 (1999)
13. Kim, J. Park, H.: Sparse nonnegative matrix factorization for clustering. Technical Report GT-CSE-08-01 (2008)
14. Mihalcea, R., Tarau, P.: Textrank: bringing order into texts. In: Proceedings of EMNLP 2004, Barcelona (2004)
15. Gao, Z.J., Song, Y., Liu, S., Wang, H., Wei, H., Chen, Y. et al.: Tracking and connecting topics via incremental hierarchical Dirichlet processes. In: 2011 IEEE 11th International Conference on Data Mining, pp. 1056–1061 (2011)
16. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. Presented at the Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France (2009)
17. Chakrabarti, D., Kumar, R., Tomkins, A.: Evolutionary clustering. Presented at the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA (2006)
18. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**, 1566 (2004)
19. Dou, W., Yu, L., Wang, X., Ma, Z., Ribarsky, W.: HierarchicalTopics: visually exploring large text collections using topic hierarchies. IEEE Trans. Vis. Comput. Graph. **19**, 2002–2011 (2013)
20. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR, vol. abs/1301.3781 (2013)
21. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: 4 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar (2014)
22. Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M.X., Qian, W. et al.: TIARA: a visual exploratory text analytic system. Presented at the Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
23. Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: visualizing thematic changes in large document collections. IEEE Trans. Vis. Comput. Graph. **8**, 9–20 (2002)
24. Cao, N., Sun, J., Lin, Y.R., Gotz, D., Liu, S., Qu, H.: FacetAtlas: multifaceted visualization for rich text corpora. IEEE Trans. Vis. Comput. Graph. **16**, 1172–1181 (2010)
25. Liu, S., Wang, X., Chen, J., Zhu, J., Guo, B.: TopicPanorama: a full picture of relevant topics. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 183–192 (2014)
26. Dou, W., Wang, X., Chang, R., Ribarsky, W.: ParallelTopics: a probabilistic approach to exploring document collections. In: 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 231–240 (2011)
27. Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., et al.: TextFlow: towards better understanding of evolving topics in text. IEEE Trans. Vis. Comput. Graph. **17**, 2412–2421 (2011)
28. Xu, P., Wu, Y., Wei, E., Peng, T.Q., Liu, S., Zhu, J.J.H., et al.: Visual analysis of topic competition on social media. IEEE Trans. Vis. Comput. Graph. **19**, 2012–2021 (2013)
29. Sun, G., Wu, Y., Liu, S., Peng, T.Q., Zhu, J.J.H., Liang, R.: EvoRiver: visual analysis of topic coopetition on social media. IEEE Trans. Vis. Comput. Graph. **20**, 1753–1762 (2014)
30. Gretarsson, B., Donovan, J.O., Bostandjiev, S., Höllerer, T., et al.: TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. Intell. Syst. Technol. **3**, 1–26 (2012)
31. Lee, H., Kihm, J., Choo, J., Stasko, J., Park, H.: iVisClustering: an interactive visual document clustering via topic modeling. Comput. Graph. Forum **31**, 1155 (2012)
32. Choo, J., Lee, C., Reddy, C.K., Park, H.: UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans. Vis. Comput. Graph. **19**, 1992–2001 (2013)

33. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America (15 November 2005)
34. El-Arini, K., Guestrin, C.: Beyond keyword search: discovering relevant scientific literature. Presented at the Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA (2011)
35. Shahaf, D., Guestrin, C., Horvitz, E.: Metro maps of science. Presented at the Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China (2012)
36. Börner, K.: Atlas of science: visualizing what we know. MIT Press, Cambridge (2010)
37. Chen, C.: CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. **57**, 359–377 (2006)
38. Pak Chung, W., Hetzler, B., Posse, C., Whiting, M., Havre, S., Cramer N. et al.: IN-SPIRE InfoVis 2004 contest entry. In: IEEE Symposium on Information Visualization, pp. r2–r2 (2004)
39. Heimerl, F., Koch, S., Bosch, H., Ertl, T.: Visual classifier training for text document retrieval. IEEE Trans. Vis. Comput. Graph. **18**, 2839–2848 (2012)
40. Lee, B., Czerwinski, M., Robertson, G., Bederson, B.B.: Understanding research trends in conferences using paperLens. Presented at the CHI '05 Extended Abstracts on Human Factors in Computing Systems, Portland, OR, USA (2005)
41. Heimerl, F., Han, Q., Koch, S., Ertl, T.: CiteRivers: visual analytics of citation patterns. IEEE Trans. Vis. Comput. Graph. **22**, 190–199 (2016)
42. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B.: Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. J Assoc. Inf. Sci. Technol 2351–2369 (2012)
43. Beck, F., Koch, S., Weiskopf, D.: Visual analysis and dissemination of scientific literature collections with SurVis. IEEE Trans. Vis. Comput. Graph. **22**, 180–189 (2016)
44. Guo, H., Laidlaw, D.H.: Topic-based exploration and embedded visualizations for research idea generation. In: IEEE Transactions on Visualization and Computer Graphics, pp. 1–1 (2018)
45. Berger, M., McDonough, K., Seversky, L.M.: cite2vec: citation-driven document exploration via word embeddings. IEEE Trans. Vis. Comput. Graph. **23**, 691–700 (2017)
46. ACM. ACM Transactions on Graphics [Online]. https://dl.acm.org/dl.cfm
47. IEEE Transactions on Visualization and Computer Graphics. ieeexplore.ieee.org/Xplore/home.jsp
48. Sheikh, Y.A., Khan, E.A., Kanade, T.: Mode-seeking by Medoidshifts. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
49. Rosner, F., Hinneburg, A., Roder, M., Nettling, M., Both, A.: Evaluating topic coherence measures (2014). arXiv:1403.6397v1
50. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. Presented at the Human Language Technologies, Los Angeles, California, 2010

Dr. Shaopeng Wu is a research fellow working in topic modelling and data visualisation at the lab of Computer Visualisation and Data Analysis, University of Bedfordshire. His main interest is natural language processing by neutral networks and deep learning. He also visualises and presents the research results on the applications by the web services. He obtained his PhD and MSc from Newcastle University in the UK, and Bachelor from Information Engineering University in China.

Youbing Zhao received his PhD in Computer Science and Engineering from Zhejiang University China in 2005. After working as an innovator in Siemens Research China for 3 years, he joined the CVDA lab of the University of Bedfordshire in 2008 and had contributed to more than 10 EU and UK EPSRC-funded projects. From 2018, he has become an associate professor in Communication University of Zhejiang China. His main research interest includes medical and information visualisation, visual analytics and intelligent computing.

Farzad Parvinzamir is a senior engineer with the Centre for Secure Information Technologies (CSIT) at Queens University Belfast. He received an MSc in computer science and a PhD in knowledge discovery and visual analytics from the University of Bedfordshire. He is currently working on distinguished security projects such as person re-identification for wide-area tracking within CSIT and also collaborating with two thriving start-ups in the area of decision analytics in sports and health care.

Nikolaos Th. Ersotelos is a lecturer in Computer Science at the University of Wolverhampton. He received a PhD degree in Computer Science from Brunel University in 2010 and an MSc (Distinction) degree in Media Production and Distribution from Lancaster University. From 2013 to 2019, he has worked at the University of Bedfordshire as a research fellow in the CVDA (Centre for Visualisation and Data Analytics). He had participated in four European Commission-funded projects. In 2016, he was involved as a Co-Investigator in a research project under a MoD contract. He has also participated in the European project "Splotch" as a researcher of Portsmouth University, in the period 2011–2012, developing an open-source software tool for visualising astrophysical cosmological simulations. From 2012–2014, he was a part-time researcher in the University of Athens (Greece), specialised in the facial expression recognition research field. His research interests include web design and development, image processing, data visualisation and immersive technologies.

Hui Wei received her degree in Computer Science and Software from Northwest University China in 1998. After working as a telecom software architect in Chinasoft International and BOCO, Beijing, China, for 10 years, she joined the CVDA lab of the University of Bedfordshire in 2008 and had contributed to more than 10 EU and UK EPSRC-funded projects. From 2016, she has become a PhD student. Her main research interest includes deep neural network, natural language processing, information visualisation, data analysis and parallel computing.

Feng Dong is a professor of Visual Computing. He was awarded a PhD from Zhejiang University. He has many research interests in computer graphics, medical visualisation and image processing. His recent work has also developed new areas in visual analytics, parallel computing and GPU, pattern recognition, image-based rendering and figure animation.